**Effat University**

**Department of Engineering**

**Factors That Impacts Accidents Severity**

Leen Zarra S19105756

Habiba Alawlaqi S19105803

Raneem Bakhashwean S19105963

College of Engineering

**Data Science**

CS 3072 2

**Supervised by:**

Dr.Zain Balfagih

Fall 2021

1. **Introduction**
2. ***Background Information & Thesis statement***

Vehicle accident mortality and damages have risen dramatically in recent decades, and they are now regarded as a serious matter of concern. According to the National Highway Traffic Safety Administration (NHTSA), there were 36,096 deaths in traffic fatalities in 2019. Every year, road accidents cost Americans hundreds of billions of dollars in economic and societal losses. And a small number of major accidents are responsible for a substantial portion of the losses. However, reducing road accidents, particularly major ones, is always a difficult task. One of the two primary ways for dealing with traffic safety issues is the proactive strategy, which focuses on avoiding potentially dangerous road situations from happening in the first place. This is due to a variety of factors. Availability of information about road conditions and potential risks can assist drivers in assessing their present position and making statistically safer driving choices. These variables might include weather fluctuations, road features, and the time of day. Road traffic accidents are the world's most significant problem. Yearly, over 1.25 deaths are caused by traffic accidents. Despite having around 54 percent of the world's cars, low- and middle-income countries account for approximately 90 percent of road deaths. Accident prediction and severity prediction are important for the successful execution of this strategy. It is possible to execute well-informed solutions and better manage financial and human resources if we can understand the patterns of how these major incidents occur and the important variables.

The project's initial goal is to find the most common factors that affect the severity of the accident and to predict the severity of the accident before it happens to minimize the number of accidents to minimize the severity of the accident. The second goal is to create a model that can reliably forecast the severity of an accident. To be more explicit, this model is expected to be able to predict the chance of a severe accident for a particular accident without any comprehensive information about the accident, such as driver qualities or vehicle type. The catastrophe might have occurred recently and details are still unavailable, or it could be one anticipated by other models.

1. ***Literature review***

Despite the advancement of traffic safety measures, highway traffic accidents remain a primary cause of mortality. The cost of fatalities and property damage caused by traffic accidents is disproportionately high in poor nations. There are several elements that contribute to traffic accidents, some of which are more important than others in determining the severity of the collision. Data mining approaches can aid in the prediction of influential accident severity variables. The primary goal of many academics' work is to identify the major elements that influence the severity of traffic accidents. Nonparametric models, linear models, and data mining approaches have all been used extensively in this type of investigation. To explain the literature on traffic accident severity prediction research, their approaches, and adopted methods have been discussed. To effectively forecast the severity of road accidents, this research employs an ensemble learning approach that blends machine learning and deep learning models. RFCNN is a suggested ensemble that combines the Random Forest (RF) and Convolutional Neural Network (CNN) models. To assess the severity of road accidents, an ensemble of regression algorithms (Voting classifier (LR+SGD)) and an ensemble of ETC and GBM are used. We also used Random Forest to identify 20 key features, which account for over half of the dataset's total accessible features. In the initial step of our experiment, we used all of the dataset's accessible attributes as input for all ensemble models. During the second part of the experiment, we fed all ensemble models the most relevant properties found by RF. Traffic accidents are the leading source of injuries, fatalities, and property devastation, and they have become a serious concern of public health and safety. Accidents also cause traffic congestion and delays. To increase the efficiency of the transportation system, accidents must be managed by analyzing relevant causes. In this work, the severity level of a traffic collision is predicted using a machine learning and deep learning model called RFCNN.

Reduced traffic accidents are a significant public safety problem. However, the majority of studies on traffic accident analysis and prediction have used small-scale datasets with limited coverage, limiting their impact and applicability; and existing large-scale datasets are either private, old, or lack critical contextual information such as environmental stimuli (weather, points-of-interest, etc.). The basic concept is to create a threshold value that optimizes the connection between POI annotations and regular expression pattern annotations. As a result, we tag the location of a collection of accident data using both regular expression patterns and OSM-based POI annotations (using a specific distance threshold). The correlation between the annotations obtained from various approaches is then measured to determine which threshold value gives the highest correlation (i.e., the best choice). A variety of data analysis and profiling on US-Accidents to extract a wide range of insights. Our analysis revealed that around 40% of accidents occurred on or near high-speed routes (highways, interstates, etc.), and approximately 32% occurred on or near local roads (streets, avenues, etc.). We also learned a lot about the relationship between accidents and time, sites of interest, and meteorological conditions. The paper's contributions: This dataset was generated over three years for the continental United States and contains over 2.25 million traffic accident records. Furthermore, the raw accident records have been supplemented with map-matching and contextual data such as weather, time of day, and places of interest. To generate a one-of-a-kind large-scale dataset of traffic events, an innovative strategy for heterogeneous data collection, cleaning, and augmentation is required. A variety of insights derived from analyses of accident hotspot locations, time, weather, and points-of-interest correlations with accident data; may be utilized directly for Urban planning, investigating weaknesses in transportation infrastructure design, traffic management, and prediction, and tailored insurance is some of the applications.

1. ***Challenges***

* Employing the huge dataset.
* The labels of this dataset are skewed.
* Lots of categorical variables.
* Determining the actual factors lead to accidents.

1. ***Data***

The data utilized within this report was collected from February 2016 to December 2019. It has been regularly collected utilizing a variety of data sources, including various APIs that give streaming traffic event data. These APIs broadcast traffic events gathered by a range of organizations, including US and state transportation authorities, law enforcement agencies, surveillance cameras, and traffic sensors embedded in road networks. This dataset currently contains over 2.9 million accident records from 49 states in the USA and 49 variables. Data cleaning procedures such as deleting null-valued rows and columns and filling null values with median and mode are conducted in the data preprocessing step. Other factors, such as "ID," "Source," and "Timezone," will not provide us with information regarding traffic accidents or be useful in estimating severity levels, so we may remove these as well. A variable name that contains "(" or ")" is risky since some functions may not be able to handle it appropriately, thus changing the variables to avoid errors. The time variables in the original dataset are in a format that is difficult to manipulate. Furthermore, if we analyze date and time as a whole, we will lose a lot of information since some patterns, such as hourly, weekly, or monthly trends, maybe concealed behind this structure. As a result, we convert time-related variables into new variables. Some variables will not be recognized as the intended type when reading the data into R. TMC, severity, and time-related factors, for example, are better treated as categorical variables than continuous variables. In addition, logistic variables should be considered categorical.

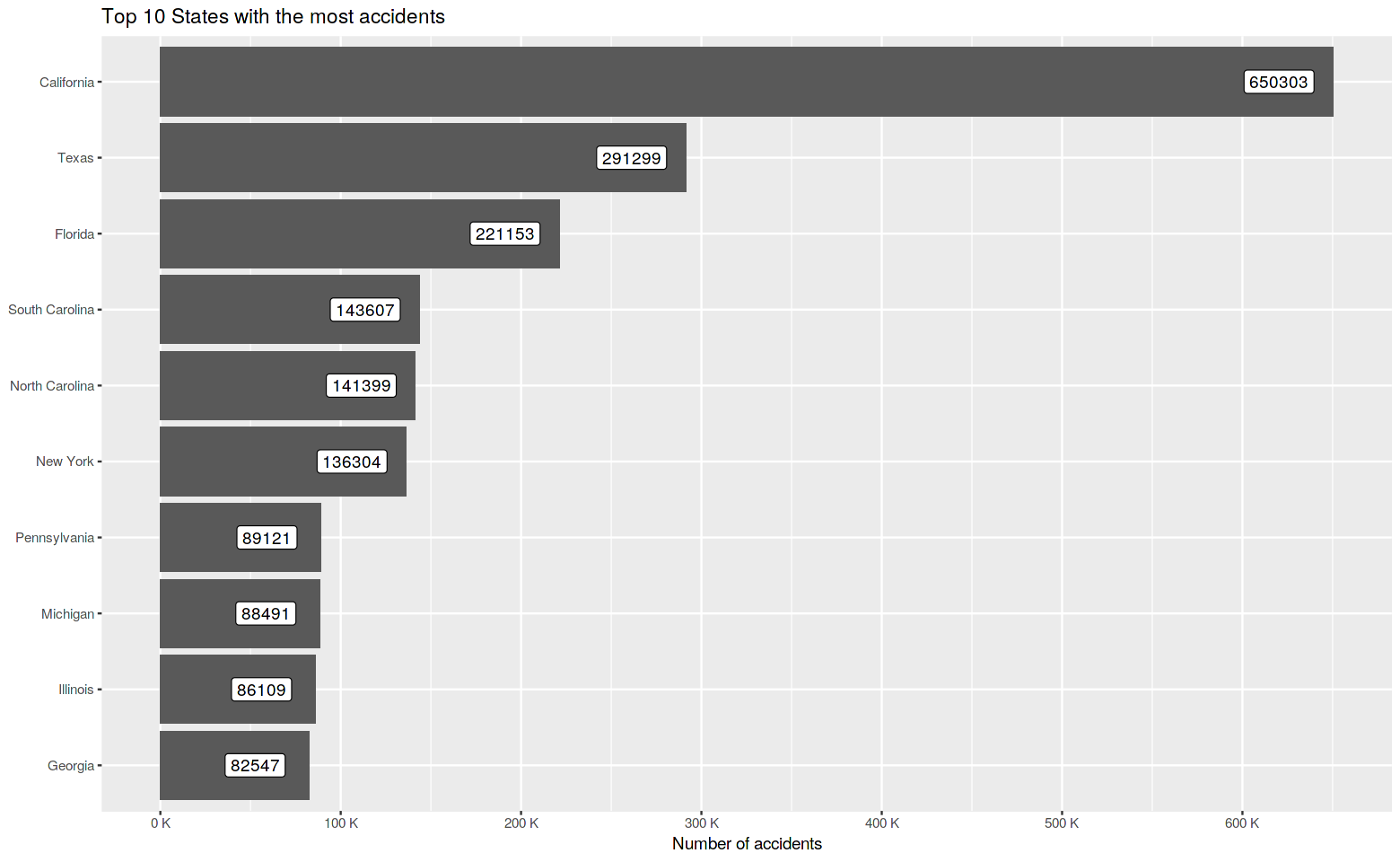
1. ***Analysis***

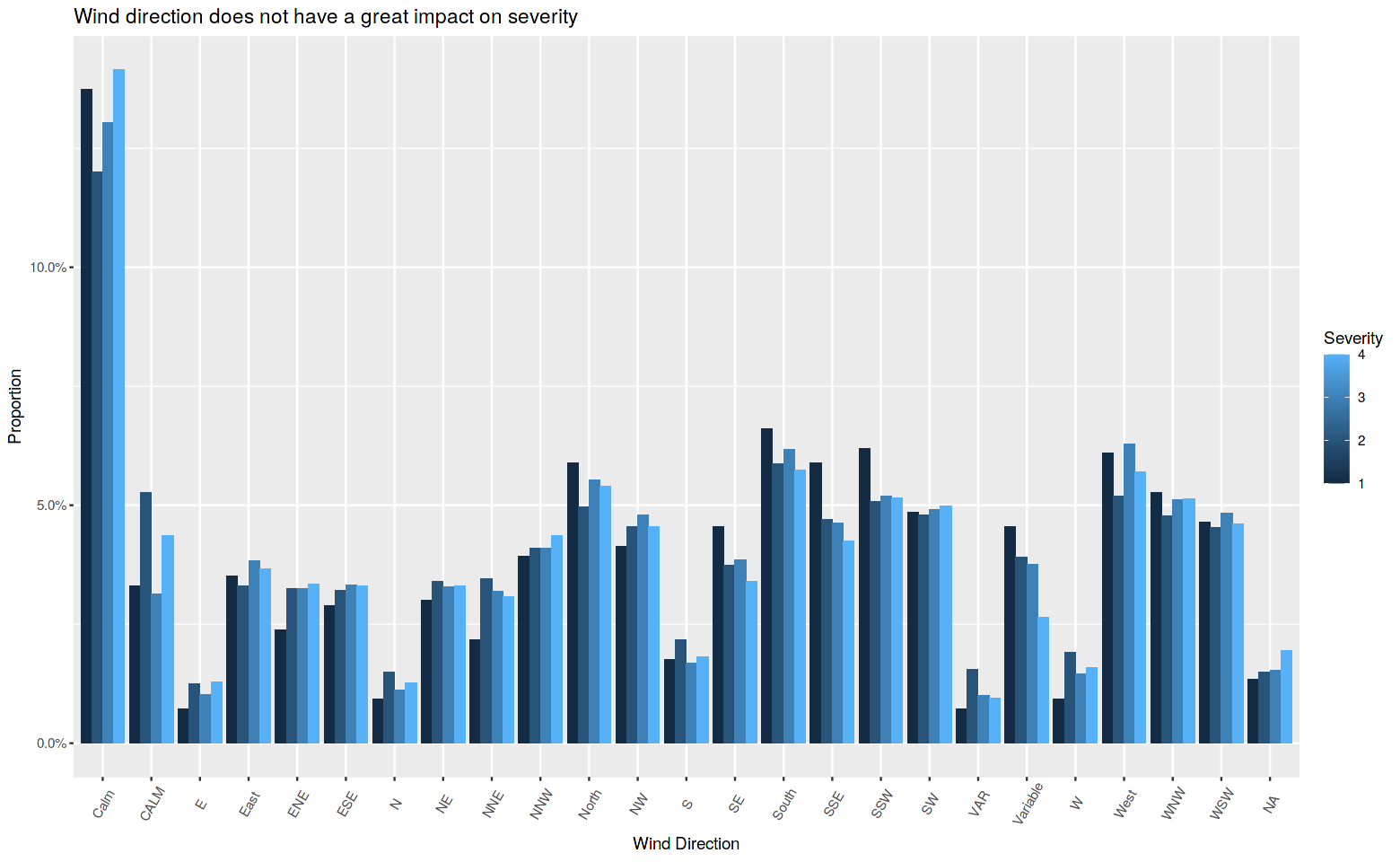
Following the data collecting and wrangling procedures, this report used the information to analyze the situation to discover the pattern between severity and other variables. This paper, on the other hand, employed geographic data visualizations to depict the accident distribution for particular factors. According to our understanding, assuming that traffic conditions do not significantly change from 2016 to 2019, the severity level distribution in each year should be similar. It shows that from 2018 to 2019, severity level 2 has a sudden increase while level 3 has a sudden decrease, which seems a little strange. Right now we still cannot give a valid explanation for this result.

To create a model that could predict the accident severity levels. Prior to conducting any analysis, the data needed to be split into training and test data sets. This is done in order that the created model can make accurate predictions and reduce the overfitting of the data. We narrow down our data to one State which is California as our target state. Some weather conditions or TMC levels only have a few records, which may cause issues when creating a model. Since the data is seriously unbalanced in different severity levels we group 4 severity levels into 2 levels, Level 1 and level 2 will be grouped as "Not Severe", and level 3 and level 4 will be grouped as "Severe". following the typical data analysis workflow by splitting the dataset into 3 sub-datasets: training(60%), validation(20%), and test(20%). Use training data to build various models, and use validation data to compare different models, and finally use test data to show the final performance.

These steps are critical to being able to produce meaningful models. This report developed 3 models to predict the accident severity levels. These are a logistic regression model, a decision tree model, and a random forest model. Each of these models offers its own strengths and weaknesses. For example, while a logistic regression model may not fit the data particularly well, it is easy to interpret. On the other hand, the random forest model may be a much better predictor, however, it is much more challenging to interpret.

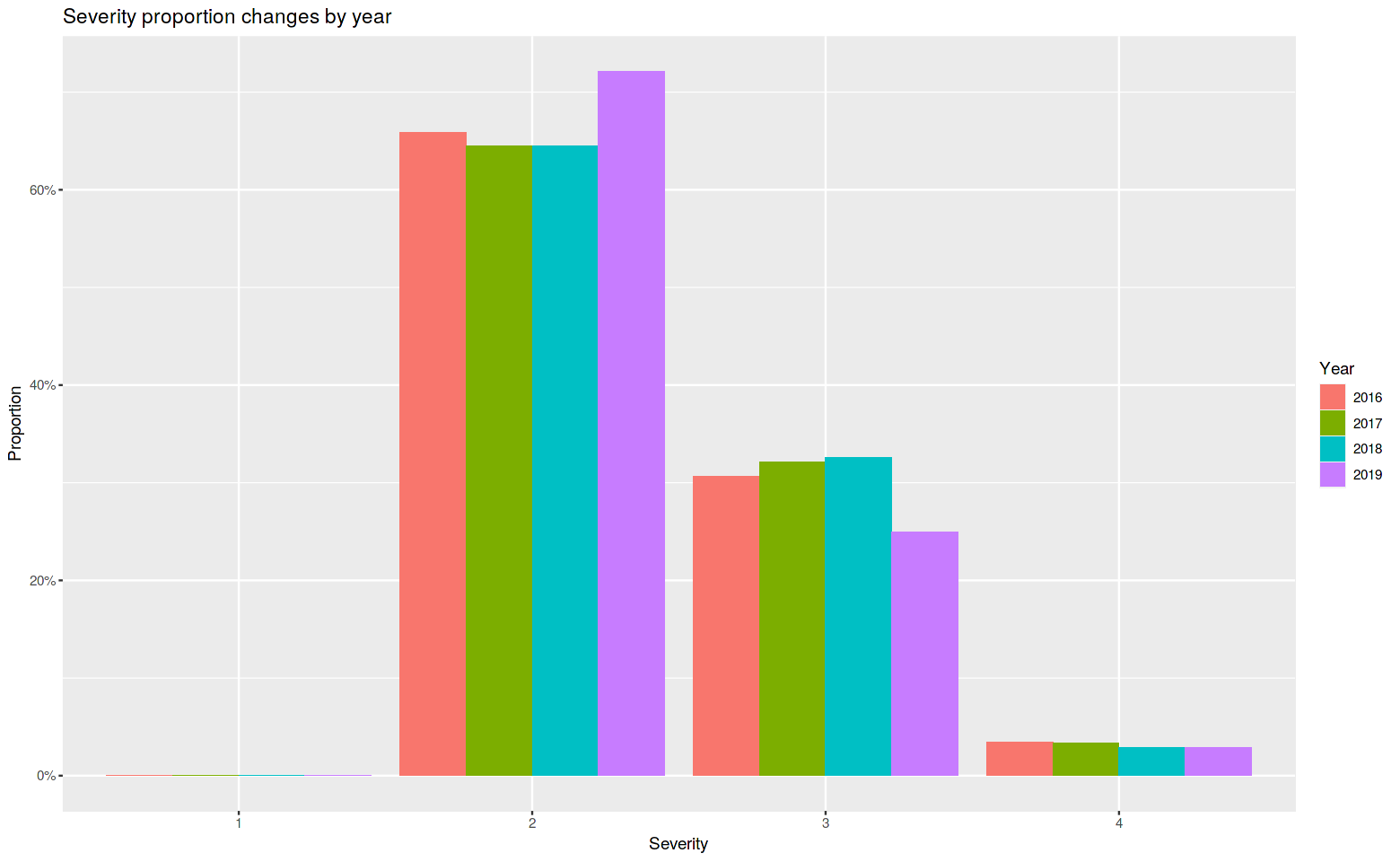
1. ***Results & Discussion***

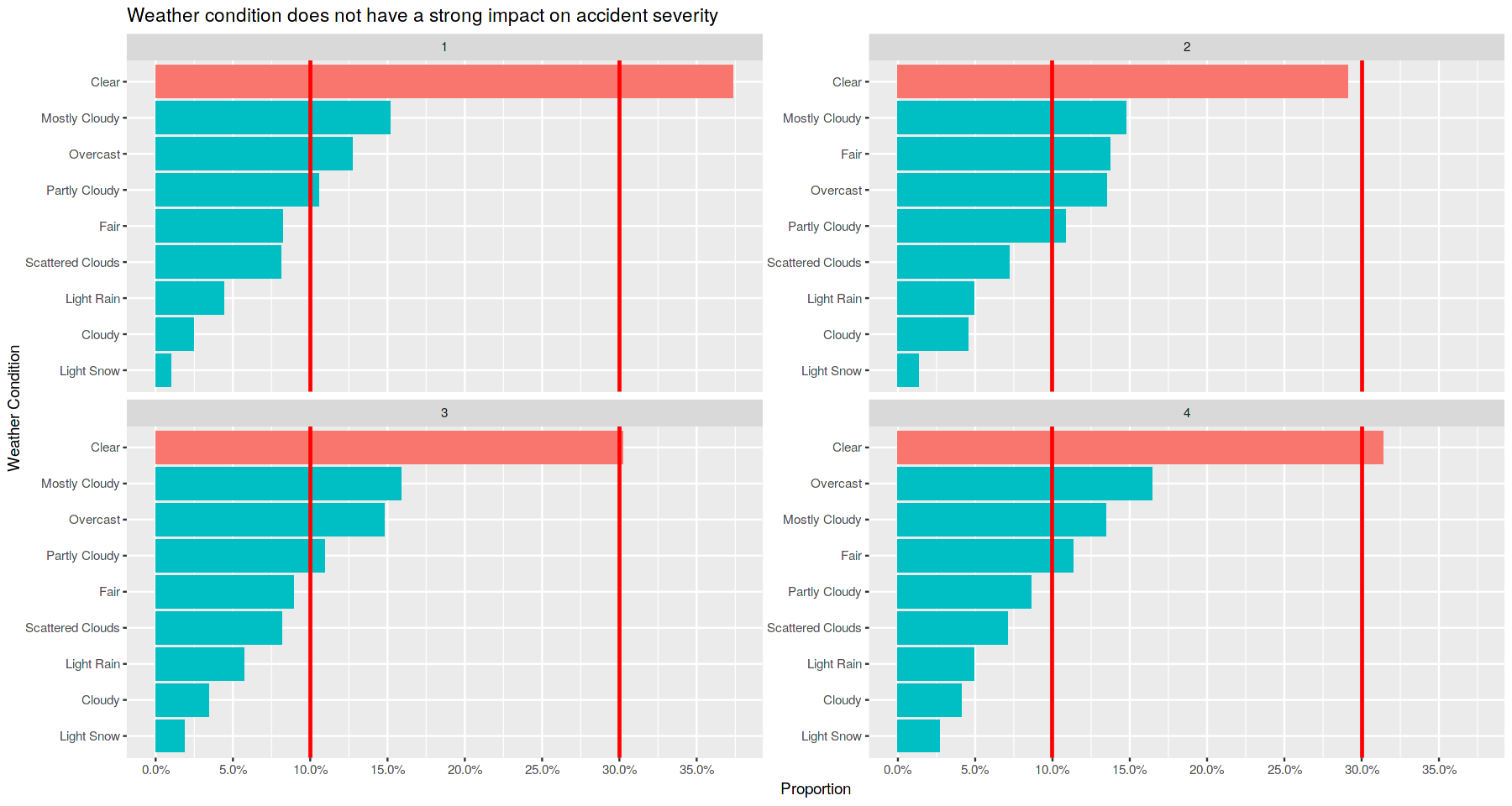




We believed "Wind\_Direction" would be a beneficial feature in our project at first. However, when we plot its distribution against each severity level, the result demonstrates that it has no influence on severity because the distributions are identical.

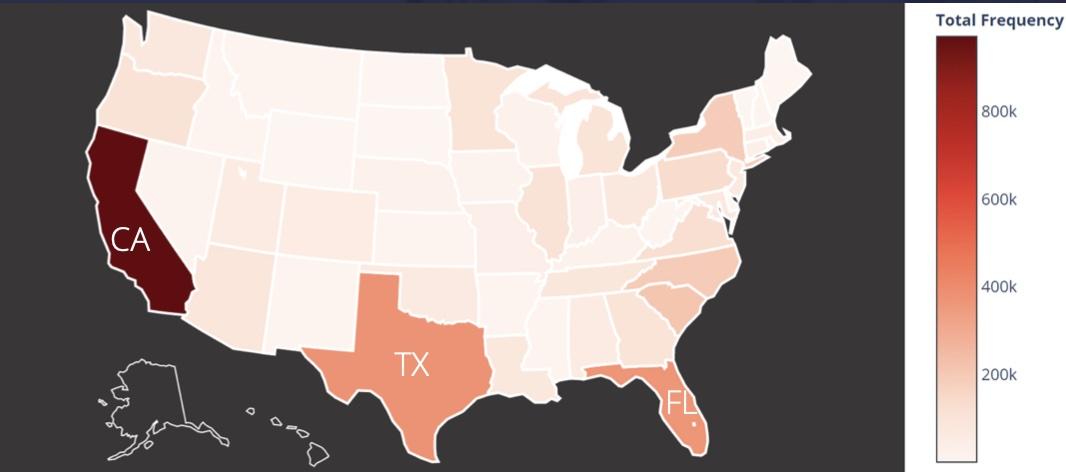
It also contains 25 levels, which adds to the model-building portion's difficulty. As a result, we've decided to abandon it.





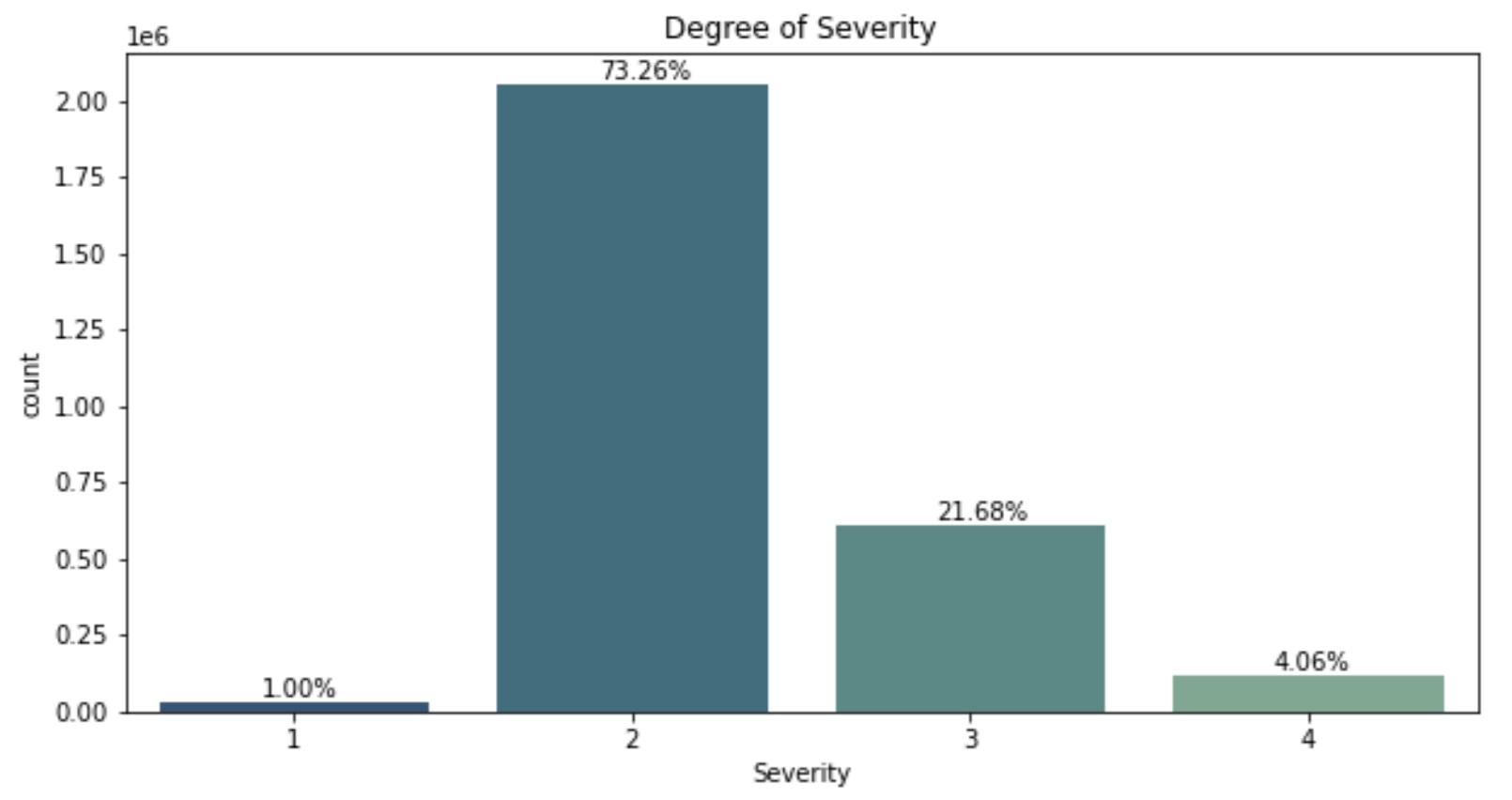
The distribution of the most typical weather situations under each severity level looks comparable at each level. Only level 1 differs in that more level 1 incident occur when the weather is clear. On clear days, we may also see that more serious accidents (levels 3 and 4) occur often. It appears that weather conditions have almost no impact on the severity of an accident.

*The state with the largest number of accidents in California, followed by Texas and Florida.*

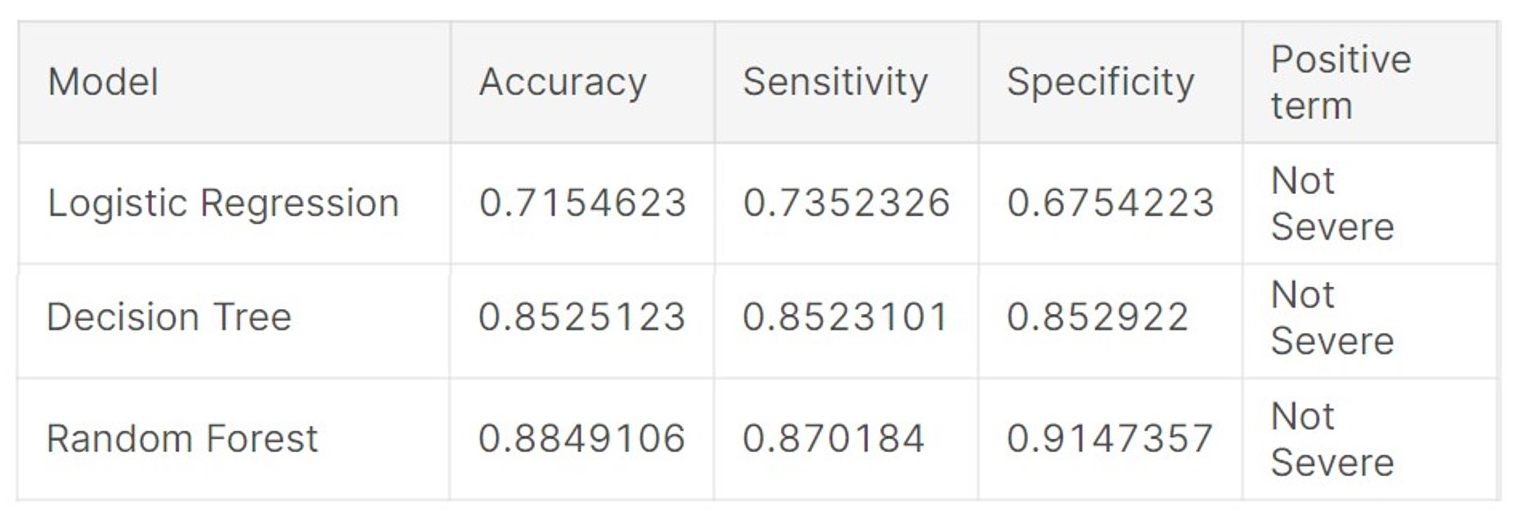


*Figure V.1 represents the US data*

The majority of the collisions occurred near a traffic light, particularly when there was a junction or a crossing.

* 
* *Figure V.2 represents the predictor variable: Severity*
* The distance of an accident is roughly proportionate to its severity, with accidents of severity 4 having the longest distance.
* The weather situation is clear in the majority of cases.
* Working days have the highest accident frequency, whereas weekends have a frequency of at least 2/3 less.
* When an accident occurs near a traffic light, it is considerably less likely to be serious, however, when it occurs at a junction, it is much more possible.

The table below shows the performance of each model:



As a result, the Random Forest model gives us the most accurate predictions. This model may be able to anticipate serious accidents in real-time using the advanced real-time traffic accident prediction solution built by the inventors of the same dataset utilized in this study.

1. ***Conclusion***

In conclusion, This paper describes US-Accidents, a one-of-a-kind, publicly available motor vehicle accident dataset, and its creation process, which includes several critical steps such as real-time traffic data collection, data integration, and multistage data augmentations using map-matching, weather, time-of-day, and points-of-interest data. To the best of our knowledge, US-Accidents is the first nationwide dataset of this magnitude, encompassing over 2.25 million traffic accident reports gathered over three years for the contiguous United States. We were able to infer a range of insights about an accident's location, timing, weather, and areas of interest from this information. Furthermore, we investigated the influence of several types of data variables on traffic accident prediction and discovered that time, traffic events, and points-of-interest have substantial significance. We want to add other publicly accessible data sources in the future for the goal of real-time traffic accident prediction.

1. ***References***

US-Accidents: A Countrywide Traffic Accident Dataset. (2021). Retrieved from <https://smoosavi.org/datasets/us_accidents>

Road traffic accidents severity factors: A review paper. (2018). Retrieved from <https://ieeexplore.ieee.org/document/8391218>

Moosavi, S., Samavatian, M., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. Retrieved from <https://arxiv.org/abs/1906.05409>

IEEE Xplore Full-Text PDF:. (2021). Retrieved from <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9536744>

Moosavi, S., Samavatian, M., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. Retrieved from <https://arxiv.org/pdf/1909.09638.pdf>